学術刊行物 情処研報 Vol.98, No.82
ISSN 0919-6072

# 情報処理学会研究報告

## 98 - NL - 127

1998年9月17日・18日

社団法人 情報処理学会

情報学基礎 51－9
自然言語処理 127－9
(1998. 9. 17)

# 多言語分散情報検索アーキテクチャに関する検討

巖寺 俊哲[1] 林 良彦[1] 菊井 玄一郎[1] 小橋 喜嗣[1] Mun-Kew Leong[2] Key-Sun Choi[3]
[1] NTT情報通信研究所
[2]Kent Ridge Digital Labs
[3]Korea Advanced Institute of Science and Technology

## 概要

本稿では、インターネット上の様々な言語の文書を提供する複数の異なるサーチエンジンをそれらの提供するコンテンツや（言語処理）機能に応じて利用することを可能にする多言語分散情報検索アーキテクチャとそこで用いられる情報検索用プロトコルについて提案する。本アーキテクチャの特徴は、メタ・サーチの採用、クロスリンガル検索のサポート、情報検索用プロトコルの使用、である。本プロトコルは、本アーキテクチャの構成要素であるメタサーチエンジンとサーチエンジンとの間で、機能やコンテンツに関する情報の共有を可能にする。これにより、メタサーチエンジンは、利用者の情報要求に応じて、複数のサーチエンジンを使い分けることが可能になる。

また、本アーキテクチャに基づいたサービスについての我々がKRDL、KAISTと行なっている共同実験プロジェクトについても紹介する。

# An Distributed Cross-Language Information Retrieval Architecture

Toshiaki IWADERA[1], Yoshihiko HAYASHI[1], Gen'ichiro KIKUI[1], Yoshitsugu OBASHI[1],
Mun-Kew Leong[2], and Key-Sun Choi[3]
[1]NTT Information and Communication Systems Laboratories
[2]Kent Ridge Digital Labs
[3]Korea Advanced Institute of Science and Technology

## Abstract

This paper proposes an architecture for distributed cross-language information retrieval and a protocol used in it. The architecture includes metasearch engines and search engines, which communicate each other by using the protocol. The protocol is designed to communicate not only a search request and its result, but also inis also introduced, and it is concerned with formation of the functionalities and the contents provided by search engines The architecture allows a user to automatically choose and exploit various search engines providing a large amount of documents in all sorts of languages in the Internet/WWW, according to his/her information need. The paper also introduces the joint project on a distributed cross-language information retrieval using the proposed architecture, involving KRDL in Singapore, KAIST in Korea and us NTT.

## 1 はじめに

近年、インターネットの発展と情報技術の進歩によって、様々な文書が容易にアクセスできるようになってきた。インターネット上の文書の大きな特徴は、それらの種類と量が膨大であることとそれらが様々な言語で書かれているということである。これらの文書を言語の違いを意識することなく、情報源として有効に活用する仕組みが必要になってきている。

多種多様な大量の文書を有効に利用するための手段として、様々かつ大規模なサーチエンジンが提供されている（たとえば、AltaVista[1]、Lycos[8]）が、これらは、主に単言語用の検索手段であり、利用者が入力した検索要求と検索対象文書の記述言語の違いを考慮していない。

検索要求と言語が異なる文書を検索する手段として、母国語で外国語の文書が検索できるクロスリンガル情報検索サービス（たとえば、TITAN[4]、CLINKS[12]）が実現されているが、これらは、特定の言語対（たとえば、日英間のみ）しか扱えない。また、検索対象データ量も比較的小さく偏っている。

これらの問題を解決するために、我々は、メタ・サーチ（たとえば、MetaCrawler[10]、SavvySearch[6]）という考え方を使って、様々な複数のサーチエンジンを統合して、利用者にインタフェースや言語の違いを意識させない多言語分散情報検索サービスを構成するためのアーキテクチャを検討している。

本稿では、インターネット上の様々な言語の文書を提供する複数の異なるサーチエンジンをそれらの提供するコンテンツや機能に応じて利用することを可能にする多言語分散情報検索アーキテクチャと検索要求とその結果の翻訳処理が介在する際の情報検索プロトコルについて提案する。

## 2 多言語分散情報検索アーキテクチャ

様々な複数のサーチエンジンを統合して、利用者にインタフェースや言語の違いを意識させない多言語分散情報検索サービスを構成するためのアーキテクチャを提案する。このアーキテクチャの特徴は、次の3点である。

1. メタ・サーチという考え方を採用している
2. クロスリンガル情報検索をサポートしている
3. 各サーチエンジンを利用するにあたって、後述する共通の検索用プロトコルを使用している

メタ・サーチという考え方を採用することで、複数のサーチエンジンを一度に検索可能になる。また、情報要求を翻訳することにより、利用者に言語の違いを意識させることなく様々な言語の情報を提供することが可能になる。さらに、共通の検索用プロトコルを用いることには、次の利点がある：

- 各サーチエンジンが提供する多様な機能を利用することが可能であり、1つのシステムで様々な機能を用意する必要がない
- 利用者の情報要求に応じて選択的に各サーチエンジンを使い分けることが可能

### 2.1 基本構成

このアーキテクチャの基本構成を図1に示す。この構成の基本構成要素は、メタサーチエンジンとサーチエンジンの2種類である。これら2種類の構成要素は、ともに、必要に応じて内部に自然言語処理機能を持つか、または、外部の同様の機能を利用する。

メタサーチエンジンは、検索対象となる情報を自らは持たず、利用者と各サーチエンジンとの間のインタフェースとなる部分である。これは、利用者の情報要求を受けるとを次のように動作する：

1. 利用者の情報要求に最適な複数のサーチエンジンを選択する。
2. 情報要求を選択したサーチエンジンが受理する言語／形式へ変換する。この過程で言語の翻訳や単語への分割、表現の統一などの自然言語処理を行なう。
3. 変換後の情報要求を用いて、選択したサーチエンジンに検索を要求する。
4. 各サーチエンジンからの検索結果を統合、編集、必要に応じて翻訳し、利用者へ提示する。

サーチエンジンは、利用者へ提供する情報を実際に格納しており、メタサーチエンジンを介して利用者から与えられる情報要求に応じて情報を提供する。各サーチエンジンは、ロボットが収集したデータや人手により作成されたデータなどを格納する1つ以上の情報ソース（以下、単にソースとよぶ）を持つ。ここでは、ソースに格納された1つのデータ単位を文書と呼ぶ。メタサーチエンジンから検索を要求されると、各サーチエンジンは、独立に、伝達された情報要求をそれが持つ機能に応じて解釈し、ソースを検索、検索結果をメタサーチエンジンへ返却する。

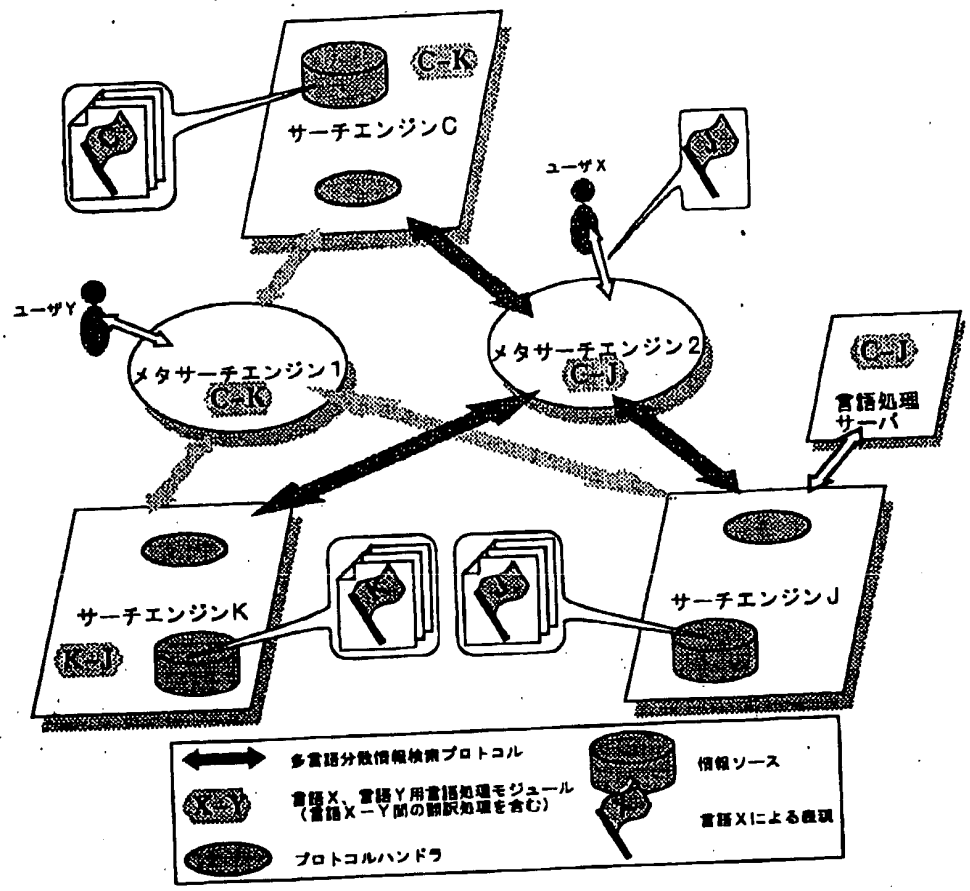


図 1: 多言語分散情報検索アーキテクチャ

本アーキテクチャでは、メタサーチエンジンとサーチエンジンとは、検索要求とその結果だけでなく、各サーチエンジンが提供する機能とコンテンツに関する情報をを共通のプロトコルを用いて必要な情報を相互に交換する。このプロトコルを用いることで、メタサーチエンジンは、利用者の情報要求に応じて、各サーチエンジンを使い分けることが可能となる。

## 2.2 多言語分散情報サーチエンジンの実現例と利点

多言語分散情報検索アーキテクチャを用いることで実現可能なサービスについて述べる。ここでは、このアーキテクチャが使用される環境を次のように想定している：

- 検索要求を記述する言語と検索対象（サーチエンジンに格納されている情報）を記述する言語が異なる場合がある
- サーチエンジン毎に提供する情報が異なる
- サーチエンジン毎に提供する機能が異なる

サーチエンジンには、単純な機能を持つものから高度な機能を提供するものまで、様々な状況が想定される。たとえば、あるサーチエンジンは、与えられた検索要求をそのまま用いて検索を実行するが、別のサーチエンジンは、クロスリンガルな検索、たとえば、与えられた検索要求をそれが検索対象とする情報に合わせて翻訳し、適切な検索式へ変換して検索するという状況である。図1では、サーチエンジンCは、言語Cで記述された情報を提供するとともに、言語Cと言語Kとの間の相互の翻訳機能を持っている。また、言語Kの情報要求が与えられると、それに対応する言語Cの情報を提供できる。同様にサーチエンジンKは、言語Kの情報を提供し、言語K－J間の翻訳機能を提供している。また、サーチエンジンJは、言語Jの情報を提供するが、その内部に翻訳機能を持っていない。しかし、外部の言語C－J間の翻訳サーバを援用することができる。さらに、メタサーチエンジン2は、その内部に言語C－J間の翻訳機能を持っている。利用者（ユー

ザX）がメタサーチエンジン２に対して、言語Ｊの情報要求を提示し、検索を要求するとこのメタサーチエンジン２は、次のように各サーチエンジンを使い分ける：

- サーチエンジンKへ：
  ユーザXから受け付けた検索要求表現をそのまま用いて検索要求する。この時、同時に、その表現の翻訳を要求する。サーチエンジンKは、検索要求を言語Kに翻訳し、検索、検索結果を言語Ｊへ翻訳しメタサーチエンジン２へ返却する。
- サーチエンジンＪへ：
  検索要求表現をそのまま用いて検索要求する。サーチエンジンＪは、検索結果をそのまま返却する。
- サーチエンジンCへ：
  メタサーチエンジン２は、言語Ｃ－Ｊ間の翻訳機能を自ら持っていることから、メタサーチエンジン２が検索要求を言語Ｃへ翻訳し、検索を要求する。サーチエンジンCは、受け付けた検索要求を用いて検索し、その検索結果をそのままメタサーチエンジン２へ返却する。メタサーチエンジン２は、ユーザXへ提示する前に検索結果を言語Ｊへ翻訳する。

以上の実現例が示すように、本アーキテクチャを用いることで次のような利点が得られる：

- 処理できる言語の拡大
  メタサーチエンジンに言語変換機能を備えることによって単言語のサーチエンジンをクロスリンガル・サーチエンジンとして利用することができる。また、既存の特定言語対のクロスリンガル・サーチエンジンを複数並列に同時に利用することにより処理可能言語を拡大することができる。たとえば、日英翻訳機能を持つサーチエンジンと日中翻訳機能を持つものを２台同時に利用することにより、利用者は、日本語の情報要求をメタサーチエンジンへ与えるだけで英語と中国語の情報の検索を行なうことができる。さらに、メタサーチエンジンが備える言語翻訳機能とサーチエンジン側が備える言語翻訳機能を同時に利用することで処理可能言語を拡大することができる（現状では、翻訳／検索精度の劣化は避けられないが）。たとえば、メタサーチエンジンが日英言語翻訳機能を備えており、あるサーチエンジンは、英中言語翻訳機能を備えていた場合、利用者は、日本語入力で中国語の文　も検索できるようになる。
- メタサーチエンジン－サーチエンジン間での機能分担
  前述した実現例中の翻訳機能のように、メタサーチエンジンとサーチエンジンのそれぞれが十分な機能を備えなくともそれぞれの提供機能を分担することで、利用者に高度な機能を提供できる。たとえば、適合フィードバックや検索結果の分類などの情報ナビゲーション機能を持たないメタサーチエンジンでも、サーチエンジン側の提供する機能を利用することで、利用者にこれらの機能を提供することができるようなる。
- 適切なサーチエンジンの選択
  多くのサーチエンジンは、大量の情報を提供しているにもかかわらず、すべての情報を網羅してはおらず、偏りが存在し、同一の検索要求に対する結果には、相違がある。これらのサーチエンジンを複数、統合的に検索することにより、情報の偏在に対処し、検索結果の網羅性を向上させることが可能となる。また、今後専門性の高いサーチエンジンが様々な分野で現れた場合、これらを利用者の情報要求に応じて、選択し、使い分けることが可能となる。
- インクリメンタルなサービス開発
  本アーキテクチャでは、インタフェースの変更等の悪影響を利用者へ与えることなくメタサーチエンジンやサーチエンジンの機能を徐々に強化していくことが可能である。たとえば、開発当初は、メタサーチエンジン、サーチエンジンとも最低限の機能（たとえば、単言語サーチエンジン）のみをまず提供し、利用者には、単言語の検索結果が提供される。しかし、サーチエンジン側に翻訳機能が備えられクロスリンガル検索対応後は、利用者には、クロスリンガル検索の結果も提供されるようになる。また、同様にこのアーキテクチャにしたがっている限りは、サーチエンジン自体を新規に追加することも容易である。

## 2.3　共通検索プロトコルに対する要件

本アーキテクチャは、メタサーチエンジンとサーチエンジンが相互に必要な情報を通知、共有し、適切に連携することが前提であり、このためには、何らかの共通検索プロトコルが必須である。ここでは、この共通検索プロトコルに要求される条件について検討する。

共通検索プロトコルに要求される機能の１つは、言語処理に関わる機能である。特に、この機能は、利用

者に言語の違いを意識することなく検索できる機能を提供する上で重要な機能である。このクロスリンガル検索機能を提供するためには、メタサーチエンジン側、または、サーチエンジン側で必然的に翻訳処理が必要となる。また、検索処理時、検索要求や検索対象文書中の語句を原型に戻したり(stemming)、表記の揺れの吸収等の表現の正規化処理やストップワードの除去等の言語処理が行なわれる。しかし、これらの言語処理技術は、言語毎に異なっており、また、同一言語でも異なるアルゴリズムや辞書等が使用されると結果が異なってしまう。これは、検索時、検索されるべきもの検索されない等の致命的な問題となる。この問題を回避するために、プロトコルに要求される機能は、言語処理機能を共有可能とする機能、あるいは、サーチエンジンで使用される言語処理機能の仕様をメタサーチエンジンへ通知する機能である。

メタサーチエンジンは、各サーチエンジンを利用するに先だって、(1)それらが提供するコンテンツ（の範囲）に関する情報と(2)機能や使用条件に関する情報サーチエンジンに関する2種類の情報を知る必要がある。これらの情報を、検索要求／結果のように検索プロセスに直接関係する情報と区別するためにメタ情報とよぶ。

コンテンツに関する情報は、メタサーチエンジンが複数のサーチエンジンの中から検索要求に対して適切なサービスを選択する場合に必要とされる情報である。サーチエンジンの選択は、サーチエンジンが出力する情報に課金していたり、情報を取得するのに時間がかかる場合、特に必須となる。メタサーチエンジンは、実際に検索を指示する前に、コンテンツ情報を利用して、検索要求に対して、各サーチエンジンを評価し、適切なサーチエンジンを選択することが可能となる。

機能や使用条件に関する情報は、メタサーチエンジンが、選択された個別のサーチエンジンを利用する際に各サーチエンジンが提供する機能を知り、利用者からの検索要求をそのサーチエンジンが受け付ける条件に合うように変換したり、あるいは、機能の適用を要求するのに必要となる情報である。これは、サーチエンジンが固有の使用条件を持っている場合などに特に重要となる。

## 3 多言語分散情報検索プロトコル

### 3.1 検索用プロトコル

検索用プロトコルとしてZ39.50[15]やSTARTS[2]が存在している。

Z39.50は、クライアント・サーバモデルに基づく情報検索のためのプロトコルである。以下の処理を行なうためのクライアントとサーバ間の通信の手順と構造を規定したものである：

- サーバによって提供されるデータベースの探索
- 探索によって同定されたデータベースレコードの検索
- ターム・リストの通覧
- 検索結果の並べ換え

さらに、アクセス制御、資源制御、拡張サービス、ヘルプ機能についても規定している。

STARTSは、簡略化されているものの、Z39.50と同様の機能を持っている。また、簡略に保つこと自体が、STARTSの提案動機の1つとなっている。さらに、Z39.50では対応していない、メタ・サーチに必要とする情報の通知機能を備えることで、メタ・サーチを行なうことを容易にしている[2]。たとえば、このような情報として、検索結果の一部として文書数とターム数を通知している。これは、複数のサーチエンジンから得られる検索結果文書のランキングを統合するとき有用となる情報である。

前述したアーキテクチャでは、クロスリンガル検索を考慮したメタ・サーチに適用可能な検索用共通プロトコルが必須である。前述したようにSTARTSは、メタ・サーチを考慮して検討されている。また、その検討の過程では、多くのサーチエンジンのベンダ（たとえば、infoseek[7]、Verity[14]）が参加しており、メタ・サーチ用プロトコルとして十分な機能を持っていると考えられる。しかし、その中では、検索要求と検索対象の記述言語が異なっている場合などのクロスリンガルな検索については考慮されていない。我々は、STARTSをベースにして、クロスリンガル検索を考慮したメタ・サーチ可能な検索用プロトコルを検討する。

### 3.2 提案するプロトコルの概要

ここでは、検索用共通プロトコル（多言語分散情報検索プロトコル、以下、プロトコルと略記する）について述べる。このプロトコルは、メタサーチエンジンとサーチエンジンの間で検索に必要とする情報の相互の交換に使用するフォーマットと手順を規定するものである。以下、このプロトコルが提供する機能について概観する。

1. 検索機能
   この機能は、メタサーチエンジンがサーチエンジ

ンへ検索を要求し、サーチエンジンから検索結果を返却するために使用するものである。検索要求時には、検索式とともにソース識別子、検索対象言語などが送られる。また、検索式中には、サーチエンジン側への検索要求表現の翻訳等の言語処理要求を記述することができ、サーチエンジン側の言語処理機能を利用することが可能となる。また、検索結果には、文書識別子や文書のタイトルだけでなく、その文書のもともとの記述言語に関する情報も返却される。

2. メタ情報取得機能
この機能は、メタサーチエンジンがサーチエンジンへメタ情報を要求し、サーチエンジンが要求されたメタ情報を返却するために使用する。返却されるメタ情報は、サーチエンジンが持つソースが格納するコンテンツに関する情報（たとえば、格納されているデータの言語、文書数、各単語の出現頻度など）や受け付ける検索式の形式条件や提供する機能（たとえば、日英翻訳機能を提供している等）である。この機能を利用することにより、適切なサーチエンジンの選択やサーチエンジン側の言語処理機能等を利用することが可能になる。

3. 文書情報取得機能
この機能は、サーチエンジンが提供する各文書に関する情報を取得するために使用する。この機能によりメタサーチエンジンは、文書毎に文書の本文やその文書への注釈情報（たとえば、その文書がどのように引用されているか等）を得ることができる。この機能を利用してメタサーチエンジンは、適合フィードバックや検索結果の分類等の情報ナビゲーション機能を利用者へ提供することが可能となる。

さらに、独立した機能としてエラー情報通知機能がある。これは、メタサーチエンジンからサーチエンジンへの要求の種類にかかわらず、サーチエンジンがその要求へ適切に応答することができない場合にその旨をメタサーチエンジンへ通知するために使用する。

本プロトコルでは、ベースにしたSTARTSに対して、クロスリンガル検索と情報ナビゲーション支援に関わる部分を中心に拡張が施されている。

クロスリンガル検索に関わる部分としては、検索機能とメタ情報取得機能が拡張されている。検索機能には、検索要求中でサーチエンジン側への言語処理要求を明示的に記述できる機能が追加されている。また、メタ情報取得機能には、メタ情報としてサーチエンジン側が提供する言語処理機能を通知する機能が追加されている。

情報ナビゲーション支援に関わる部分としては、検索機能が拡張されているとともに、STARTSでは、全く考慮されていなかった文書情報取得機能が追加されている。検索機能には、検索要求中でサーチエンジン側へ情報ナビゲーション支援（たとえば、Query Expanstionや検索結果の分類等）を要求できる機能が追加されている。また、文書取得機能が追加されていることで、各文書毎に詳細な情報得られ、この情報を用いることでメタサーチエンジン側でナビゲーション支援機能を提供することが可能になっている。

### 3.3 検索要求例

本プロトコルでは、メタサーチエンジンとサーチエンジンとの間でやり取りされる情報は、属性と属性値のペアのリストとして表現される。図2に、HARVEST[3]で使用されているSOIF形式に従って記述した検索要求の一例を示す。（これはあくまでも一例であり、他の形式（たとえば、XML）に従って記述しても良い。）

この検索要求は、次の条件を記述したものである。

- titleフィールドに「ワイン工場の歴史」を、また、body-of-textフィールドに「カリフォルニア」を含む文書の検索を要求する。また、検索結果として複数の文書が得られた場合は、body-of-textフィールドに「ロゼ」を多く含む順に出力すること。但し、実際の検索には、入力された表現（「ワイン工場の歴史」、「カリフォルニア」、「ロゼ」）からキーワードのみを抽出し、英訳して使用すること。
- 検索対象とする情報ソースは、Source-1。
- 検索結果として出力されるのは、各検索結果文書内のtitleフィールドとauthorフィールドのみでよい。
- 検索結果文書は、USドメインから収集されたもので、内容は、英語で書かれていること。
- 検索結果文書のtitleは、日本語へ翻訳して出力すること。

## 4 共同実験プロジェクト

ここでは、我々が、KRDL（Kent Ridge Digital Labs、シンガポール）、KAIST（Korea Advanced Institute of Science and Technology、韓国）とともに進めている共同実験プロジェクトについて紹介する。

```
@Query{
  Version{9}: CL-MS 0.0
  FilterExpressti n{103}:
        ((title Translate Tokenize ''ワイン工場の歴史'')
        and (body-of-text Translate Tokenize ''カリフォルニア''))
  RankingExpression{40}:
        (body-of-text Translate Tokenize ''ロゼ'')
  DefaultLanguage{5}: ja-JP
  Sources{8}: Source-1
  AnswerFields{12}: title author
  DocumentDomain{2}: US
  DocumentLanguage{2}: en
  DocumentTranslationLevel{5}: title
  DocumentTranslationTargetLanguage{2}: ja
}
```

図 2: 検索要求例

このプロジェクトの目的は、多言語分散情報検索サービスの実現であり、実際に、日本、シンガポール、韓国の３ヶ所に検索サーバを設置し、利用者がそれらを１つのメタサーチエンジンから検索できる環境を構築、公開することを目指している。

このサービスで対象とする言語は、日本語、中国語、韓国語、英語である。利用者は、ブラウザを介して提供される１つのメタサーチエンジンへのインタフェースを用いて、上記言語のいずれかを利用して検索を要求することができる。この検索表現は、各言語へ翻訳され、翻訳された検索表現にマッチした文書が検索結果として利用者へ提示される。

このプロジェクトのメイン・テーマは、前述した多言語分散情報検索プロトコルの開発であり、共同で実現するのは、プロトコルの定義である。全体システムの実現に必要なメタサーチエンジン、検索サーバ、各言語処理モジュールは、それぞれ独自に開発、作成する予定である。

## 5 議論

多種多様な大量の文書を有効に利用するための手段として、様々なサーチエンジンが提供されている。しかし、利用者が既存のサーチエンジンを利用しようとした場合、以下の点が問題となる：(1) 検索を始めるにあたって、利用者は、各自が持つ情報要求を各サーチエンジンが受け付ける形式で表現する必要がある（検索質問表現の記述／入力の問題）。また、(2) すべてのサーチエンジンが検索結果として同様の情報を返却するのではないため、複数のサーチエンジンを使用したり、使い分けたりする必要がある（提供情報の問題）。さらに、(3) 利用者の望む情報が異言語で書かれている可能性がある場合は、情報要求を利用者自ら異言語へ翻訳し表現する必要がある（言葉の壁の問題）。以下、これらの問題点とここで提案している本アーキテクチャでの対処法について考察する。

検索質問表現の記述／入力の問題に対しては、既存のいくつかのサーチエンジンでは、利用者が適切に質問表現を記述するのを支援したり、大量の検索結果を閲覧するのを支援する機能（情報ナビゲーション支援機能）を提供している。情報ナビゲーションを支援する機能として、適合フィードバック [9] や検索結果の編集提示（たとえば、検索結果のクラスタリング表示 [5]、関連タームの提案 [13]）、GUI による検索条件の直接操作 [11] などがある。このような機能を実現するためには、多くのサーチエンジンで検索結果として提示される文書のタイトルや参照情報（たとえば、URL）だけでなく、文書の本文等のより詳細な情報を必要とする。本プロトコルでは、文書情報取得機能により詳細な文書情報を提供することができ、この情報を利用してメタサーチエンジンは、情報ナビゲーション支援機能を実現することができる。

サーチエンジンの使い分けの問題は、メタ・サーチという考え方を導入することで解決される。この考え方は、利用者からみて複数のサーチエンジンをあたかも１つのように見せ、１回の操作でこれら複数のサーチエンジンを検索可能にするものである。本アーキテクチャでは、メタ・サーチの考え方を用いており、利用者がサーチエンジンを使い分けるのを代行することが可能である。さらに、各サーチエンジンからメタ情報としてそれが提供しているコンテンツに関する情報を得ており、実際に各サーチエンジンに検索要求を発行する前に、利用者の情報要求に対応して、適切なサーチエンジンを選択することが可能である。このため、すべてのサーチエンジンへ検索を要求する場合に比較

して効率的に検索を行なうことができる。

利用者が言語の壁を意識することなく検索を行なうためには、利用者が運用可能である言語（X言語）を用いて異言語（Y言語）の情報を検索可能であり、その検索結果が適切であるかを判定できるようにその内容のすべて、あるいは、その要約等をX言語で提示可能にすることが必要である。さらに、利用者が検索結果の情報を読みこなすためには、翻訳などの支援が必要である。本アーキテクチャでは、検索要求中に検索要求表現の翻訳をサーチエンジンへ要求することができる。また、サーチエンジンは、メタ情報の一部としてそれがどのような翻訳機能などの言語処理機能を提供しているかを公開することができる。メタサーチエンジンは、この情報を利用して翻訳機能を持つサーチエンジンをあらかじめ選択し、利用することができ、利用者に言語の違いを意識させることなく、情報を提供することが可能となる。

## 6 おわりに

本稿では、多言語分散情報検索アーキテクチャとこの実装に必須であるプロトコルについて提案した。メタ・サーチには、依然として技術的側面、サービス的側面ともに様々な問題が存在する。クロスリンガル検索についても同様である。我々の多言語分散情報検索サービスに関する共同実験は、これらの有用性を試す最初の試金石となると考えられる。今後は、このサービスを公開し、ここで提案したアーキテクチャ、プロトコルを検証、評価していく予定である。

本稿では、プロトコルに関しては、その概要のみについて述べた。このプロトコルの詳細な仕様、および、ここで提案したアーキテクチャを我々が実装した実験サービスは、http://titan.isl.ntt.co.jp/dclir/ で公開する予定である。

## 参考文献


[1] AltaVista. http://www.altavista.digital.com.

[2] Luis Gravano, Chen-Chuan K. Chang, Héctor García-Molina, and Andreas Paepcke. Starts: Stanford proposal for internet meta-searching. Technical report, Stanford University, 1997.

[3] Accesible at http://harvest.transarc.com/.

[4] Y. Hayashi, G. Kikui, and S. Susaki. TITAN: A cr ss-linguistic search engine for the www. In *AAAI Spring Symposiun on Cross-Language Text and Speech Retrieval*, 1997. Accesible at http://titan.navi.ntt.co.jp.

[5] M. Hearst. Interfaces for searching the web. *Scientific American*, pp. 68-72, March 1997.

[6] Adele Howe and Daniel Dreilinger. SavvySearch: A meta-search engine that learns which search engines to query. *AI Magazine*, Vol. 18, No. 2,, 1997.
Accesible at http://guaraldi.cs.colostate.edu:2000/for

[7] Infoseek. http://www.infoseek.com.

[8] Lycos. http://www.lycos.com.

[9] Gerard Salton. *Automatic Text Processing The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[10] Erik Selberg and Oren Etzioni. Multi-service search and comparison using the MetaCrawler. In *Proceedings of the 4th International World Wide Web Conference*, 1995.
Accesible at http://www.metacrawler.com/.

[11] 鷲崎, 林, 菊井. WWW上の情報探索システムにおけるインタラクティブイン タフェース. インタラクティブシステムとソフトウェアIV. 日本ソフトウェア科学会, 近代科学社, 1996.

[12] 鈴木雅実, 井ノ上直己, 橘本和夫. 多言語情報検索における利用者支援について — 主要キーワードの対訳付与に関する検討 —. 情報処理学会自然言語処理研究会資料 NL122-11, 1997. Accesible at http://mlc.kddvw.kcom.or.jp/CLINKS/.

[13] 鈴木雅実, 井ノ上直己, 橘本和夫. クロスリンガル情報検索結果の閲覧支援のための主要キーワード対訳表示の効果. 情報処理学会自然言語処理研究会資料 NL126-14, 1998.

[14] Verity. http://www.verity.com.

[15] *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995).*

学術刊行物 情処研報 Vol.98, No.82
ISSN 0919-6072

# 情報処理学会研究報告

## 98 - NL - 127

1998年9月17日・18日

社団法人 情報処理学会

情報学基礎 51－9
自然言語処理 127－9
(1998. 9. 17)

# 多言語分散情報検索アーキテクチャに関する検討

巖寺 俊哲[1] 林 良彦[1] 菊井 玄一郎[1] 小橋 喜嗣[1] Mun-Kew Leong[2] Key-Sun Choi[3]

[1] NTT情報通信研究所

[2]Kent Ridge Digital Labs

[3]Korea Advanced Institute of Science and Technology

## 概要

本稿では、インターネット上の様々な言語の文書を提供する複数の異なるサーチエンジンをそれらの提供するコンテンツや（言語処理）機能に応じて利用することを可能にする多言語分散情報検索アーキテクチャとそこで用いられる情報検索用プロトコルについて提案する。本アーキテクチャの特徴は、メタ・サーチの採用、クロスリンガル検索のサポート、情報検索用プロトコルの使用、である。本プロトコルは、本アーキテクチャの構成要素であるメタサーチエンジンとサーチエンジンとの間で、機能やコンテンツに関する情報の共有を可能にする。これにより、メタサーチエンジンは、利用者の情報要求に応じて、複数のサーチエンジンを使い分けることが可能になる。

また、本アーキテクチャに基づいたサービスについての我々がKRDL、KAISTと行なっている共同実験プロジェクトについても紹介する。

# An Distributed Cross-Language Information Retrieval Architecture

Toshiaki IWADERA[1], Yoshihiko HAYASHI[1], Gen'ichiro KIKUI[1], Yoshitsugu OBASHI[1], Mun-Kew Leong[2], and Key-Sun Choi[3]

[1]NTT Information and Communication Systems Laboratories

[2]Kent Ridge Digital Labs

[3]Korea Advanced Institute of Science and Technology

## Abstract

This paper proposes an architecture for distributed cross-language information retrieval and a protocol used in it. The architecture includes metasearch engines and search engines, which communicate each other by using the protocol. The protocol is designed to communicate not only a search request and its result, but also inis also introduced, and it is concerned with formation of the functionalities and the contents provided by search engines The architecture allows a user to automatically choose and exploit various search engines providing a large amount of documents in all sorts of languages in the Internet/WWW, according to his/her information need. The paper also introduces the joint project on a distributed cross-language information retrieval using the proposed architecture, involving KRDL in Singapore, KAIST in Korea and us NTT.

## 1 はじめに

近年、インターネットの発展と情報技術の進歩によって、様々な文　が容易にアクセスできるようになってきた。インターネット上の文書の大きな特徴は、それらの種類と量が膨大であることとそれらが様々な言語で書かれているということである。これらの文書を言語の違いを意識することなく、情報源として有効に活用する仕組みが必要になってきている。

多種多様な大量の文書を有効に利用するための手段として、様々かつ大規模なサーチエンジンが提供されている（たとえば、AltaVista[1]、Lycos[8]）が、これらは、主に単言語用の検索手段であり、利用者が入力した検索要求と検索対象文書の記述言語の違いを考慮していない。

検索要求と言語が異なる文書を検索する手段として、母国語で外国語の文書が検索できるクロスリンガル情報検索サービス（たとえば、TITAN[4]、CLINKS[12]）が実現されているが、これらは、特定の言語対（たとえば、日英間のみ）しか扱えない。また、検索対象データ量も比較的小さく偏っている。

これらの問題を解決するために、我々は、メタ・サーチ（たとえば、MetaCrawler[10]、SavvySearch[6]）という考え方を使って、様々な複数のサーチエンジンを統合して、利用者にインタフェースや言語の違いを意識させない多言語分散情報検索サービスを構成するためのアーキテクチャを検討している。

本稿では、インターネット上の様々な言語の文書を提供する複数の異なるサーチエンジンをそれらの提供するコンテンツや機能に応じて利用することを可能にする多言語分散情報検索アーキテクチャと検索要求とその結果の翻訳処理が介在する際の情報検索プロトコルについて提案する。

## 2 多言語分散情報検索アーキテクチャ

様々な複数のサーチエンジンを統合して、利用者にインタフェースや言語の違いを意識させない多言語分散情報検索サービスを構成するためのアーキテクチャを提案する。このアーキテクチャの特徴は、次の3点である。

1. メタ・サーチという考え方を採用している
2. クロスリンガル情報検索をサポートしている
3. 各サーチエンジンを利用するにあたって、後述する共通の検索用プロトコルを使用している

メタ・サーチという考え方を採用することで、複数のサーチエンジンを一度に検索可能になる。また、情報要求を翻訳することにより、利用者に言語の違いを意識させることなく様々な言語の情報を提供することが可能になる。さらに、共通の検索用プロトコルを用いることには、次の利点がある：

- 各サーチエンジンが提供する多様な機能を利用することが可能であり、1つのシステムで様々な機能を用意する必要がない
- 利用者の情報要求に応じて選択的に各サーチエンジンを使い分けることが可能

### 2.1 基本構成

このアーキテクチャの基本構成を図1に示す。この構成の基本構成要素は、メタサーチエンジンとサーチエンジンの2種類である。これら2種類の構成要素は、ともに、必要に応じて内部に自然言語処理機能を持つか、または、外部の同様の機能を利用する。

メタサーチエンジンは、検索対象となる情報を自らは持たず、利用者と各サーチエンジンとの間のインタフェースとなる部分である。これは、利用者の情報要求を受けるとを次のように動作する：

1. 利用者の情報要求に最適な複数のサーチエンジンを選択する。
2. 情報要求を選択したサーチエンジンが受理する言語／形式へ変換する。この過程で言語の翻訳や単語への分割、表現の統一などの自然言語処理を行なう。
3. 変換後の情報要求を用いて、選択したサーチエンジンに検索を要求する。
4. 各サーチエンジンからの検索結果を統合、編集、必要に応じて翻訳し、利用者へ提示する。

サーチエンジンは、利用者へ提供する情報を実際に格納しており、メタサーチエンジンを介して利用者から与えられる情報要求に応じて情報を提供する。各サーチエンジンは、ロボットが収集したデータや人手により作成されたデータなどを格納する1つ以上の情報ソース（以下、単にソースとよぶ）を持つ。ここでは、ソースに格納された1つのデータ単位を文　と呼ぶ。メタサーチエンジンから検索を要求されると、各サーチエンジンは、独立に、伝達された情報要求をそれが持つ機能に応じて解釈し、ソースを検索、検索結果をメタサーチエンジンへ返却する。

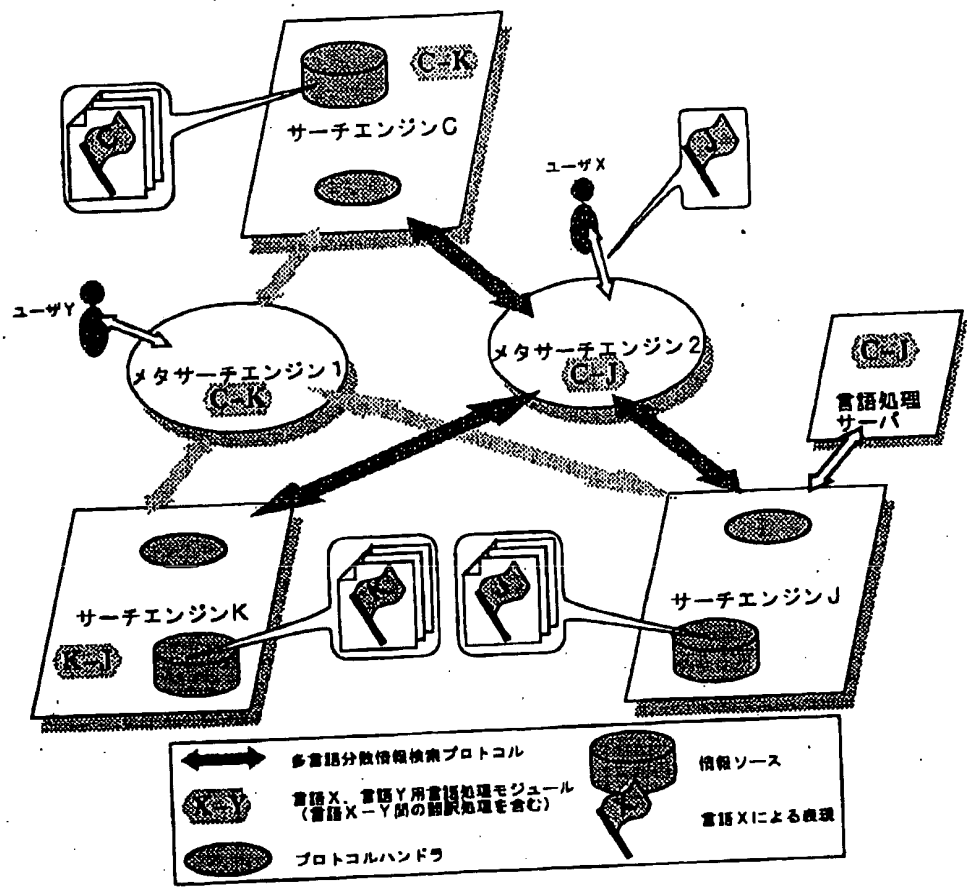


図 1: 多言語分散情報検索アーキテクチャ

本アーキテクチャでは、メタサーチエンジンとサーチエンジンとは、検索要求とその結果だけでなく、各サーチエンジンが提供する機能とコンテンツに関する情報をを共通のプロトコルを用いて必要な情報を相互に交換する。このプロトコルを用いることで、メタサーチエンジンは、利用者の情報要求に応じて、各サーチエンジンを使い分けることが可能となる。

## 2.2 多言語分散情報サーチエンジンの実現例と利点

多言語分散情報検索アーキテクチャを用いることで実現可能なサービスについて述べる。ここでは、このアーキテクチャが使用される環境を次のように想定している：

- 検索要求を記述する言語と検索対象（サーチエンジンに格納されている情報）を記述する言語が異なる場合がある
- サーチエンジン毎に提供する情報が異なる
- サーチエンジン毎に提供する機能が異なる

サーチエンジンには、単純な機能を持つものから高度な機能を提供するものまで、様々な状況が想定される。たとえば、あるサーチエンジンは、与えられた検索要求をそのまま用いて検索を実行するが、別のサーチエンジンは、クロスリンガルな検索、たとえば、与えられた検索要求をそれが検索対象とする情報に合わせて翻訳し、適切な検索式へ変換して検索するという状況である。図1では、サーチエンジンCは、言語Cで記述された情報を提供するとともに、言語Cと言語Kとの間の相互の翻訳機能を持っている。また、言語Kの情報要求が与えられると、それに対応する言語Cの情報を提供できる。同様にサーチエンジンKは、言語Kの情報を提供し、言語K－J間の翻訳機能を提供している。また、サーチエンジンJは、言語Jの情報を提供するが、その内部に翻訳機能を持っていない。しかし、外部の言語C－J間の翻訳サーバを援用することができる。さらに、メタサーチエンジン2は、その内部に言語C－J間の翻訳機能を持っている。利用者（ユー

ザX）がメタサーチエンジン２に対して、言語Ｊの情報要求を提示し、検索を要求するとこのメタサーチエンジン２は、次のように各サーチエンジンを使い分ける：

- サーチエンジンKへ：
  ユーザXから受け付けた検索要求表現をそのまま用いて検索要求する。この時、同時に、その表現の翻訳を要求する。サーチエンジンKは、検索要求を言語Kに翻訳し、検索、検索結果を言語Ｊへ翻訳しメタサーチエンジン２へ返却する。
- サーチエンジンＪへ：
  検索要求表現をそのまま用いて検索要求する。サーチエンジンＪは、検索結果をそのまま返却する。
- サーチエンジンCへ：
  メタサーチエンジン２は、言語Ｃ－Ｊ間の翻訳機能を自ら持っていることから、メタサーチエンジン２が検索要求を言語Ｃへ翻訳し、検索を要求する。サーチエンジンCは、受け付けた検索要求を用いて検索し、その検索結果をそのままメタサーチエンジン２へ返却する。メタサーチエンジン２は、ユーザXへ提示する前に検索結果を言語Ｊへ翻訳する。

以上の実現例が示すように、本アーキテクチャを用いることで次のような利点が得られる：

- 処理できる言語の拡大
  メタサーチエンジンに言語変換機能を備えることによって単言語のサーチエンジンをクロスリンガル・サーチエンジンとして利用することができる。また、既存の特定言語対のクロスリンガル・サーチエンジンを複数並列に同時に利用することにより処理可能言語を拡大することができる。たとえば、日英翻訳機能を持つサーチエンジンと日中翻訳機能を持つものを２台同時に利用することにより、利用者は、日本語の情報要求をメタサーチエンジンへ与えるだけで英語と中国語の情報の検索を行なうことができる。さらに、メタサーチエンジンが備える言語翻訳機能とサーチエンジン側が備える言語翻訳機能を同時に利用することで処理可能言語を拡大することができる（現状では、翻訳／検索精度の劣化は避けられないが）。たとえば、メタサーチエンジンが日英言語翻訳機能を備えており、あるサーチエンジンは、英中言語翻訳機能を備えていた　合、利用者は、日本語入力で中国語の文　も検索できるようになる。
- メタサーチエンジンーサーチエンジン間での機能分担
  前述した実現例中の翻訳機能のように、メタサーチエンジンとサーチエンジンのそれぞれが十分な機能を備えなくともそれぞれの提供機能を分担することで、利用者に高度な機能を提供できる。たとえば、適合フィードバックや検索結果の分類などの情報ナビゲーション機能を持たないメタサーチエンジンでも、サーチエンジン側の提供する機能を利用することで、利用者にこれらの機能を提供することができるようなる。
- 適切なサーチエンジンの選択
  多くのサーチエンジンは、大量の情報を提供しているにもかかわらず、すべての情報を網羅してはおらず、偏りが存在し、同一の検索要求に対する結果には、相違がある。これらのサーチエンジンを複数、統合的に検索することにより、情報の偏在に対処し、検索結果の網羅性を向上させることが可能となる。また、今後専門性の高いサーチエンジンが様々な分野で現れた場合、これらを利用者の情報要求に応じて、選択し、使い分けることが可能となる。
- インクリメンタルなサービス開発
  本アーキテクチャでは、インタフェースの変更等の悪影響を利用者へ与えることなくメタサーチエンジンやサーチエンジンの機能を徐々に強化していくことが可能である。たとえば、開発当初は、メタサーチエンジン、サーチエンジンとも最低限の機能（たとえば、単言語サーチエンジン）のみをまず提供し、利用者には、単言語の検索結果が提供される。しかし、サーチエンジン側に翻訳機能が備えられクロスリンガル検索対応後は、利用者には、クロスリンガル検索の結果も提供されるようになる。また、同様にこのアーキテクチャにしたがっている限りは、サーチエンジン自体を新規に追加することも容易である。

## 2.3　共通検索プロトコルに対する要件

本アーキテクチャは、メタサーチエンジンとサーチエンジンが相互に必要な情報を通知、共有し、適切に連携することが前提であり、このためには、何らかの共通検索プロトコルが必須である。ここでは、この共通検索プロトコルに要求される条件について検討する。

共通検索プロトコルに要求される機能の１つは、言語処理に関わる機能である。特に、この機能は、利用

者に言語の違いを意識することなく検索できる機能を提供する上で重要な機能である。このクロスリンガル検索機能を提供するためには、メタサーチエンジン側、または、サーチエンジン側で必然的に翻訳処理が必要となる。また、検索処理時、検索要求や検索対象文書中の語句を原型に戻したり(stemming)、表記の揺れの吸収等の表現の正規化処理やストップワードの除去等の言語処理が行なわれる。しかし、これらの言語処理技術は、言語毎に異なっており、また、同一言語でも異なるアルゴリズムや辞書等が使用されると結果が異なってしまう。これは、検索時、検索されるべきもの検索されない等の致命的な問題となる。この問題を回避するために、プロトコルに要求される機能は、言語処理機能を共有可能とする機能、あるいは、サーチエンジンで使用される言語処理機能の仕様をメタサーチエンジンへ通知する機能である。

メタサーチエンジンは、各サーチエンジンを利用するに先だって、(1)それらが提供するコンテンツ（の範囲）に関する情報と(2)機能や使用条件に関する情報サーチエンジンに関する2種類の情報を知る必要がある。これらの情報を、検索要求／結果のように検索プロセスに直接関係する情報と区別するためにメタ情報とよぶ。

コンテンツに関する情報は、メタサーチエンジンが複数のサーチエンジンの中から検索要求に対して適切なサービスを選択する場合に必要とされる情報である。サーチエンジンの選択は、サーチエンジンが出力する情報に課金していたり、情報を取得するのに時間がかかる場合、特に必須となる。メタサーチエンジンは、実際に検索を指示する前に、コンテンツ情報を利用して、検索要求に対して、各サーチエンジンを評価し、適切なサーチエンジンを選択することが可能となる。

機能や使用条件に関する情報は、メタサーチエンジンが、選択された個別のサーチエンジンを利用する際に各サーチエンジンが提供する機能を知り、利用者からの検索要求をそのサーチエンジンが受け付ける条件に合うように変換したり、あるいは、機能の適用を要求するのに必要となる情報である。これは、サーチエンジンが固有の使用条件を持っている場合などに特に重要となる。

## 3 多言語分散情報検索プロトコル

### 3.1 検索用プロトコル

検索用プロトコルとしてZ39.50[15]やSTARTS[2]が存在している。

Z39.50は、クライアント・サーバモデルに基づく情報検索のためのプロトコルである。以下の処理を行なうためのクライアントとサーバ間の通信の手順と構造を規定したものである：

- サーバによって提供されるデータベースの探索
- 探索によって同定されたデータベースレコードの検索
- ターム・リストの通覧
- 検索結果の並べ換え

さらに、アクセス制御、資源制御、拡張サービス、ヘルプ機能についても規定している。

STARTSは、簡略化されているものの、Z39.50と同様の機能を持っている。また、簡略に保つこと自体が、STARTSの提案動機の1つとなっている。さらに、Z39.50では対応していない、メタ・サーチに必要とする情報の通知機能を備えることで、メタ・サーチを行なうことを容易にしている[2]。たとえば、このような情報として、検索結果の一部として文書数とターム数を通知している。これは、複数のサーチエンジンから得られる検索結果文書のランキングを統合するとき有用となる情報である。

前述したアーキテクチャでは、クロスリンガル検索を考慮したメタ・サーチに適用可能な検索用共通プロトコルが必須である。前述したようにSTARTSは、メタ・サーチを考慮して検討されている。また、その検討の過程では、多くのサーチエンジンのベンダ（たとえば、infoseek[7]、Verity[14]）が参加しており、メタ・サーチ用プロトコルとして十分な機能を持っていると考えられる。しかし、その中では、検索要求と検索対象の記述言語が異なっている場合などのクロスリンガルな検索については考慮されていない。我々は、STARTSをベースにして、クロスリンガル検索を考慮したメタ・サーチ可能な検索用プロトコルを検討する。

### 3.2 提案するプロトコルの概要

ここでは、検索用共通プロトコル（多言語分散情報検索プロトコル、以下、プロトコルと略記する）について述べる。このプロトコルは、メタサーチエンジンとサーチエンジンの間で検索に必要とする情報の相互の交換に使用するフォーマットと手順を規定するものである。以下、このプロトコルが提供する機能について概観する。

1. 検索機能

   この機能は、メタサーチエンジンがサーチエンジ

ンへ検索を要求し、サーチエンジンから検索結果を返却するために使用するものである。検索要求時には、検索式とともにソース識別子、検索対象言語などが送られる。また、検索式中には、サーチエンジン側への検索要求表現の翻訳等の言語処理要求を記述することができ、サーチエンジン側の言語処理機能を利用することが可能となる。また、検索結果には、文書識別子や文書のタイトルだけでなく、その文書のもともとの記述言語に関する情報も返却される。

2. メタ情報取得機能
この機能は、メタサーチエンジンがサーチエンジンへメタ情報を要求し、サーチエンジンが要求されたメタ情報を返却するために使用する。返却されるメタ情報は、サーチエンジンが持つソースが格納するコンテンツに関する情報（たとえば、格納されているデータの言語、文書数、各単語の出現頻度など）や受け付ける検索式の形式条件や提供する機能（たとえば、日英翻訳機能を提供している等）である。この機能を利用することにより、適切なサーチエンジンの選択やサーチエンジン側の言語処理機能等を利用することが可能になる。

3. 文書情報取得機能
この機能は、サーチエンジンが提供する各文書に関する情報を取得するために使用する。この機能によりメタサーチエンジンは、文書毎に文書の本文やその文書への注釈情報（たとえば、その文書がどのように引用されているか等）を得ることができる。この機能を利用してメタサーチエンジンは、適合フィードバックや検索結果の分類等の情報ナビゲーション機能を利用者へ提供することが可能となる。

さらに、独立した機能としてエラー情報通知機能がある。これは、メタサーチエンジンからサーチエンジンへの要求の種類にかかわらず、サーチエンジンがその要求へ適切に応答することができない場合にその旨をメタサーチエンジンへ通知するために使用する。

本プロトコルでは、ベースにしたSTARTSに対して、クロスリンガル検索と情報ナビゲーション支援に関わる部分を中心に拡張が施されている。

クロスリンガル検索に関わる部分としては、検索機能とメタ情報取得機能が拡張されている。検索機能には、検索要求中でサーチエンジン側への言語処理要求を明示的に記述できる機能が追加されている。また、メタ情報取得機能には、メタ情報としてサーチエンジン側が提供する言語処理機能を通知する機能が追加されている。

情報ナビゲーション支援に関わる部分としては、検索機能が拡張されているとともに、STARTSでは、全く考慮されていなかった文書情報取得機能が追加されている。検索機能には、検索要求中でサーチエンジン側へ情報ナビゲーション支援（たとえば、Query Expanstionや検索結果の分類等）を要求できる機能が追加されている。また、文書取得機能が追加されていることで、各文書毎に詳細な情報得られ、この情報を用いることでメタサーチエンジン側でナビゲーション支援機能を提供することが可能になっている。

### 3.3 検索要求例

本プロトコルでは、メタサーチエンジンとサーチエンジンとの間でやり取りされる情報は、属性と属性値のペアのリストとして表現される。図2に、HARVEST[3]で使用されているSOIF形式に従って記述した検索要求の一例を示す。（これはあくまでも一例であり、他の形式（たとえば、XML）に従って記述しても良い。）

この検索要求は、次の条件を記述したものである。

- titleフィールドに「ワイン工場の歴史」を、また、body-of-textフィールドに「カリフォルニア」を含む文書の検索を要求する。また、検索結果として複数の文書が得られた場合は、body-of-textフィールドに「ロゼ」を多く含む順に出力すること。但し、実際の検索には、入力された表現（「ワイン工場の歴史」、「カリフォルニア」、「ロゼ」）からキーワードのみを抽出し、英訳して使用すること。
- 検索対象とする情報ソースは、Source-1。
- 検索結果として出力されるのは、各検索結果文書内のtitleフィールドとauthorフィールドのみでよい。
- 検索結果文書は、USドメインから収集されたもので、内容は、英語で書かれていること。
- 検索結果文書のtitleは、日本語へ翻訳して出力すること。

## 4 共同実験プロジェクト

ここでは、我々が、KRDL (Kent Ridge Digital Labs、シンガポール)、KAIST (Korea Advanced Institute of Science and Technology、韓国)とともに進めている共同実験プロジェクトについて紹介する。

```
@Query{
  Version{9}: CL-MS 0.0
  FilterExpressti n{103}:
        ((title Translate Tokenize ''ワイン工場の歴史'')
        and (body-of-text Translate Tokenize ''カリフォルニア''))
  RankingExpression{40}:
        (body-of-text Translate Tokenize ''ロゼ'')
  DefaultLanguage{5}: ja-JP
  Sources{8}: Source-1
  AnswerFields{12}: title author
  DocumentDomain{2}: US
  DocumentLanguage{2}: en
  DocumentTranslationLevel{5}: title
  DocumentTranslationTargetLanguage{2}: ja
}
```

図 2: 検索要求例

このプロジェクトの目的は、多言語分散情報検索サービスの実現であり、実際に、日本、シンガポール、韓国の３ヶ所に検索サーバを設置し、利用者がそれらを１つのメタサーチエンジンから検索できる環境を構築、公開することを目指している。

このサービスで対象とする言語は、日本語、中国語、国語、英語である。利用者は、ブラウザを介して提供される１つのメタサーチエンジンへのインタフェースを用いて、上記言語のいずれかを利用して検索を要求することができる。この検索表現は、各言語へ翻訳され、翻訳された検索表現にマッチした文書が検索結果として利用者へ提示される。

このプロジェクトのメイン・テーマは、前述した多言語分散情報検索プロトコルの開発であり、共同で実現するのは、プロトコルの定義である。全体システムの実現に必要なメタサーチエンジン、検索サーバ、各言語処理モジュールは、それぞれ独自に開発、作成する予定である。

## 5 議論

多種多様な大量の文書を有効に利用するための手段として、様々なサーチエンジンが提供されている。しかし、利用者が既存のサーチエンジンを利用しようとした場合、以下の点が問題となる：(1) 検索を始めるにあたって、利用者は、各自が持つ情報要求を各サーチエンジンが受け付ける形式で表現する必要がある（検索質問表現の記述／入力の問題）。また、(2) すべてのサーチエンジンが検索結果として同様の情報を返却するのではないため、複数のサーチエンジンを使用したり、使い分けたりする必要がある（提供情報の問題）。さらに、(3) 利用者の望む情報が異言語で かれている可能性がある場合は、情報要求を利用者自ら異言語へ翻訳し表現する必要がある（言葉の壁の問題）。以下、これらの問題点とここで提案している本アーキテクチャでの対処法について考察する。

検索質問表現の記述／入力の問題に対しては、既存のいくつかのサーチエンジンでは、利用者が適切に質問表現を記述するのを支援したり、大量の検索結果を閲覧するのを支援する機能（情報ナビゲーション支援機能）を提供している。情報ナビゲーションを支援する機能として、適合フィードバック [9] や検索結果の編集提示（たとえば、検索結果のクラスタリング表示 [5]、関連タームの提案 [13]）、GUI による検索条件の直接操作 [11] などがある。このような機能を実現するためには、多くのサーチエンジンで検索結果として提示される文書のタイトルや参照情報（たとえば、URL）だけでなく、文書の本文等のより詳細な情報を必要とする。本プロトコルでは、文書情報取得機能により詳細な文書情報を提供することができ、この情報を利用してメタサーチエンジンは、情報ナビゲーション支援機能を実現することができる。

サーチエンジンの使い分けの問題は、メタ・サーチという考え方を導入することで解決される。この考え方は、利用者からみて複数のサーチエンジンをあたかも１つのように見せ、１回の操作でこれら複数のサーチエンジンを検索可能にするものである。本アーキテクチャでは、メタ・サーチの考え方を用いており、利用者がサーチエンジンを使い分けるのを代行することが可能である。さらに、各サーチエンジンからメタ情報としてそれが提供しているコンテンツに関する情報を得ており、実際に各サーチエンジンに検索要求を発行する前に、利用者の情報要求に対応して、適切なサーチエンジンを選択することが可能である。このため、すべてのサーチエンジンへ検索を要求する 合に比較

して効率的に検索を行なうことができる。

利用者が言語の壁を意識することなく検索を行なうためには、利用者が運用可能である言語（X言語）を用いて異言語（Y言語）の情報を検索可能であり、その検索結果が適切であるかを判定できるようにその内容のすべて、あるいは、その要約等をX言語で提示可能にすることが必要である。さらに、利用者が検索結果の情報を読みこなすためには、翻訳などの支援が必要である。本アーキテクチャでは、検索要求中に検索要求表現の翻訳をサーチエンジンへ要求することができる。また、サーチエンジンは、メタ情報の一部としてそれがどのような翻訳機能などの言語処理機能を提供しているかを公開することができる。メタサーチエンジンは、この情報を利用して翻訳機能を持つサーチエンジンをあらかじめ選択し、利用することができ、利用者に言語の違いを意識させることなく、情報を提供することが可能となる。

## 6 おわりに

本稿では、多言語分散情報検索アーキテクチャとこの実装に必須であるプロトコルについて提案した。メタ・サーチには、依然として技術的側面、サービス的側面ともに様々な問題が存在する。クロスリンガル検索についても同様である。我々の多言語分散情報検索サービスに関する共同実験は、これらの有用性を試す最初の試金石となると考えられる。今後は、このサービスを公開し、ここで提案したアーキテクチャ、プロトコルを検証、評価していく予定である。

本稿では、プロトコルに関しては、その概要のみについて述べた。このプロトコルの詳細な仕様、および、ここで提案したアーキテクチャを我々が実装した実験サービスは、http://titan.isl.ntt.co.jp/dclir/ で公開する予定である。

## 参考文献


[1] AltaVista. http://www.altavista.digital.com.

[2] Luis Gravano, Chen-Chuan K. Chang, Héctor García-Molina, and Andreas Paepcke. Starts: Stanford proposal for internet meta-searching. Technical report, Stanford University, 1997.

[3] Accesible at http://harvest.transarc.com/.

[4] Y. Hayashi, G. Kikui, and S. Susaki. TITAN: A cross-linguistic search engine for the www. In *AAAI Spring Symposiun on Cross-Language Text and Speech Retrieval*, 1997. Accesible at http://titan.navi.ntt.c jp.

[5] M. Hearst. Interfaces for searching the web. *Scientific American*, pp. 68-72, March 1997.

[6] Adele Howe and Daniel Dreilinger. SavvySearch: A meta-search engine that learns which search engines to query. *AI Magazine*, Vol. 18, No. 2,, 1997.
Accesible at http://guaraldi.cs.colostate.edu:2000/form

[7] Infoseek. http://www.infoseek.com.

[8] Lycos. http://www.lycos.com.

[9] Gerard Salton. *Automatic Text Processing The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[10] Erik Selberg and Oren Etzioni. Multi-service search and comparison using the MetaCrawler. In *Proceedings of the 4th International World Wide Web Conference*, 1995.
Accesible at http://www.metacrawler.com/.

[11] 鷲崎, 林, 菊井. WWW 上の情報探索システムにおけるインタラクティブイン タフェース. インタラクティブシステムとソフトウェア IV. 日本ソフトウェア科学会, 近代科学社, 1996.

[12] 鈴木雅実, 井ノ上直己, 橋本和夫. 多言語情報検索における利用者支援について — 主要キーワードの対訳付与に関する検討 —. 情報処理学会自然言語処理研究会資料 NL122-11, 1997. Accesible at http://mlc.kddvw.kcom.or.jp/CLINKS/.

[13] 鈴木雅実, 井ノ上直己, 橋本和夫. クロスリンガル情報検索結果の閲覧支援のための主要キーワード対訳表示の効果. 情報処理学会自然言語処理研究会資料 NL126-14, 1998.

[14] Verity. http://www.verity.com.

[15] *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995)*.